Home Work 3

Harshith Bondada
M10724854

1. Read <tombstone.csv> into R. Use response variable = Marble Tombstone Mean Surface Recession Rate, and covariate = Mean SO2 concentrations over a 100 year period. Description: Marble Tombstone Mean Surface Recession Rates and Mean SO2 concentrations over a 100-year period.

We are reading the data of the tombstone into home_work3 and taking the response variable y as the Surface Recession Rate and covariate x as the SO2 concentration.

```
home_work3=read.csv("tombstone.csv")

attach(home_work3)
y=Marble.Tombstone.Mean.Surface.Recession.Rate..mm.100years.
x=Modelled.100.Year.Mean.SO2.Concentration..ug.m..3.
model1=lm(y~x, data=home_work3)
plot(y~x)
abline(model1, lwd=2)
summary(model1)
summary(model1)$r.square
```

In the below output screenshot, we can interpret that the data is successfully read into the homework3 and a model1 has been generated.

```
  cannot open file 'tombstone(1).csv': No such file or directory
> home_work3=read.csv("tombstone.csv")
> attach(home_work3)
> y=Marble.Tombstone.Mean.Surface.Recession.Rate..mm.100years.
> x=Modelled.100.Year.Mean.SO2.Concentration..ug.m..3.
> model1=lm(y~x, data=home_work3)
> plot(y~x)
> abline(model1, lwd=2)
> summary(model1)
```

2. Obtain R^2, explain what it means.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

The higher the R-squared, the better the model fits your data.

```
summary(model1)
summary(model1)$r.square
```

The summary of the model is shown in the below screenshot.

We get the R2 to be around 81%. This means that the 81% of the response variable can be explained by the linear regression model.

This means that the more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line.

Here, this means that the regression line is 81% good fit for the data.

```
> summary(model1)

Call:
lm(formula = y ~ x, data = home_work2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72384 -0.19138  0.06136  0.13320  0.69412 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.3229959  0.1521958   2.122   0.0472 *  
x           0.0085933  0.0009499   9.046 2.58e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.365 on 19 degrees of freedom
Multiple R-squared:  0.8116,	Adjusted R-squared:  0.8017 
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08

> summary(model1)$r.square
[1] 0.8115724
```

3. Perform the following hypothesis testing and interval estimation using lm() and other related R functions.

3.1. Perform t tests, obtain t statistics and p values, interpret the results, make a conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

Null Hypothesis:

If there is a significant **linear** relationship between the independent variable X and the dependent variable Y, the slope will not equal zero. The **null hypothesis** states that the slope is equal to zero, and the **alternative hypothesis** states that the slope is not equal to zero.

We can know the values of the t test, it's statistics and p value by using the summary function.

```
summary(model1)
summary(model1)$coeff[,3]
```

Here, we find the t value for the slope to be 9.046, which means that the slope is 9.046 standard error's above 0, which is greater than 2. Hence, we reject the null hypothesis.

Similarly, for the intercept, we find it's t value to be 2.122, which means that the intercept is 2.122 standard error's above 0, which is greater than 2. Hence, we reject the null hypothesis.

We also get to know the P value from the summary function.

The P value for the intercept is 0.0472 with one * significant code which is close 0.01 according to significant codes. Since, $P<0.05$ we reject the Null Hypothesis.

Similarly, with the slope it is significantly close to 0. Hence, we reject the null hypothesis.

Harshith Bondada
M10724854

```
> summary(model1)

Call:
lm(formula = y ~ x, data = home_work3)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72384 -0.19138  0.06136  0.13320  0.69412 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.3229959  0.1521958   2.122   0.0472 *  
x           0.0085933  0.0009499   9.046 2.58e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.365 on 19 degrees of freedom
Multiple R-squared:  0.8116,    Adjusted R-squared:  0.8017 
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08

> summary(model1)$coeff[,3]
(Intercept)           x 
   2.122239    9.046242 
>
```

3.2. Perform ANOVA test (F test), obtain F statistic and p value, interpret the results, make conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

Null hypothesis: The fit of the intercept-only model and your model are equal.

We can find the F-statistic and the p value using the summary function.

```
summary(model1)
```

The values of the F-statistic and p value are found at the bottom of the summary.

We find the ratio for the variances is 81.3 which is large F value.

When there is a large F value, the corresponding P value is small.

Here, the p value is less than 0.05, we reject the Null hypothesis.

Since, the ratio is large, this means that the good portion of the variances can be decomposed to the linear regression model. i.e., we have a very significant slope.

```
Residual standard error: 0.365 on 19 degrees of freedom
Multiple R-squared:  0.8116,    Adjusted R-squared:  0.8017 
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08
```

3.3. Compute confidence interval for coefficients, fitted values (mean response), interpret the meanings of these quantities.

We can compute the **confidence interval** for the coefficients using the **confint( )** function.

```
confint(model1)
```

This interval contains the true parameter with 95% probability.

Interval estimate [0.004446349, 0.64154545]. This contains the true intercept 95% probability.

Slope estimate [0.006605098, 0.01058157]. This contains the true slope with 95% probability.

```
> confint(model1)
                  2.5 %     97.5 %
(Intercept) 0.004446349 0.64154545
x           0.006605098 0.01058157
> 
```

We can compute the **confidence interval** for the fitted values using the **predict()** function.

I am first creating data frame for y and x.

And then generating the head of the data frame for explaining it easily.

You can see in the below screenshot that I have used the predict function with two arguments. The first one takes our generated model and the second one takes the type of interval. Since, we want confidence, we mention it's a confidence interval.

```
d=data.frame(y,x)
head(d)
predict.lm(model1, interval="confidence")
```

Harshith Bondada
M10724854

```
> head(d)
     y  x
1 0.27 12
2 0.14 20
3 0.33 20
4 0.81 46
5 0.84 48
6 1.08 92
```

This means that for a observation of y and x

The point estimate is between the lwr and upr values.

Let me explain, with a simple example. Let us take the first observation.

This means that for y=0.27 with regressor x= 12, the point y^ estimate is 0.426 in the interval range [0.127, 0.7245] with 95% probability.

```
> predict.lm(model1, interval="confidence")
         fit       lwr       upr
1  0.4261159 0.1276356 0.7245962
2  0.4948626 0.2094375 0.7802876
3  0.4948626 0.2094375 0.7802876
4  0.7182892 0.4729586 0.9636199
5  0.7354759 0.4930475 0.9779043
6  1.1135825 0.9248239 1.3023412
7  1.1049892 0.9152900 1.2946885
8  1.1307692 0.9438424 1.3176960
9  1.1995159 1.0192244 1.3798074
10 1.3284159 1.1572428 1.4995889
11 1.3713825 1.2021867 1.5405784
12 1.5432492 1.3761806 1.7103178
13 1.5432492 1.3761806 1.7103178
14 1.8526092 1.6666152 2.0386032
15 1.8697959 1.6820050 2.0575867
16 2.0158825 1.8103314 2.2214336
17 2.2479025 2.0069821 2.4888230
18 2.3338359 2.0781916 2.5894801
19 2.3768025 2.1135420 2.6400631
20 2.4197692 2.1487417 2.6907967
21 3.0986425 2.6921265 3.5051585
>
```

Harshith Bondada
M10724854

3.4. Plot data points, the regression line, the confidence interval for fitted values (to show that the interval is wider on both sides and narrow in the center).

We can obtain the confidence intervals using the lines function.

We first create a new variable and specify it's range.

Since, our range is beyond 300 on the x-axis, we specify the range in newx to be around (0, 350).

```
newx<-seq(0,350)
conf<-predict(model1,newdata=data.frame(x=newx),interval = c("confidence"),level = 0.95,type="response")
plot(x,y,pch=20)
model1 <- lm(y ~ x, data=home_work3)
abline(model1,col="blue")
lines(newx,conf[,2],col="red",lty=100)
lines(newx,conf[,3],col="red",lty=100)
```

The below screenshot describes the regression line, and the confidence interval

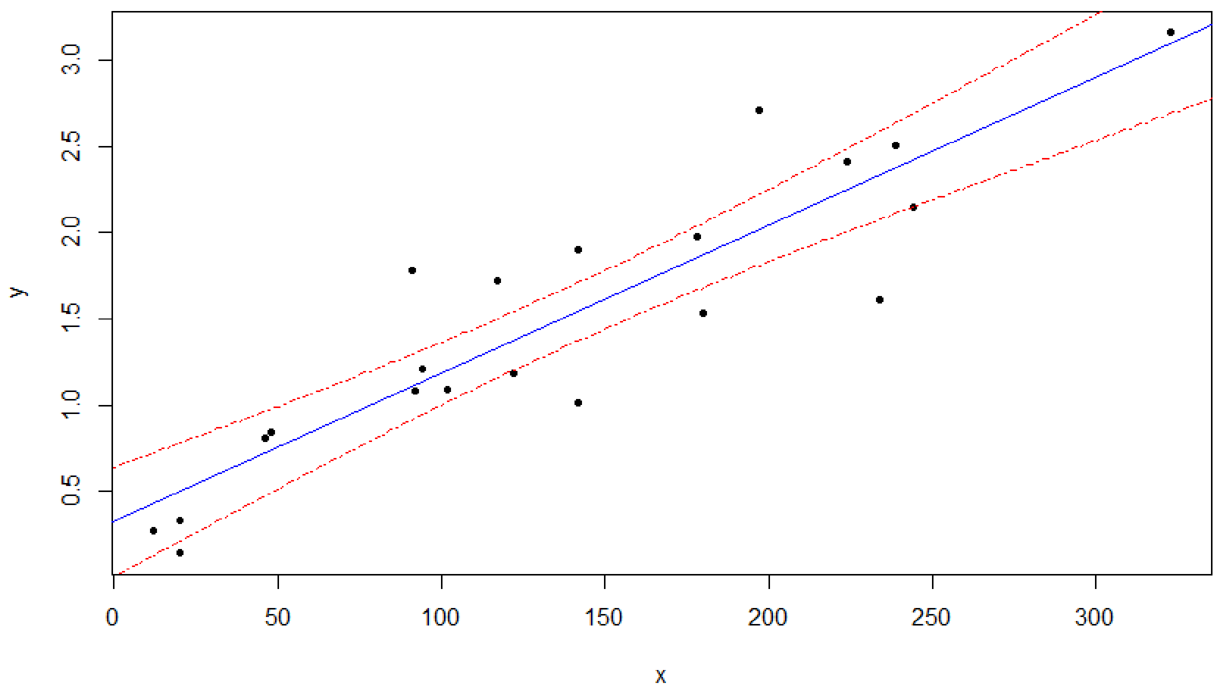
Lines for the values.

The blue line is the regression line. The other two red lines are the upper bound line and the lower bound line.

The interval estimation for the mean response of the first observation is [0,0.510].

In other words, [0, 0.510] contains the true mean response with 95% probability.

Harshith Bondada
M10724854



Harshith Bondada
M10724854

4. Using the output from summary (), suppose we want to test for null hypothesis of $H_0: \beta_1 = 0.01$ against the alternative hypothesis $H_1: \beta_1 \neq 0.01$, what do you conclude? Reject or not reject? Explain why.

Here, we do not reject the null hypothesis for B_1=0.01.

We can clearly see that the t test say's that the B_1 is -1.481 SD error's above the null hypothesis. When taken the absolute value we get +1.481 which is less than 2.

Hence, we do not reject the null hypothesis.

The p value is also greater than 0.05. This implies that we do not reject the null hypothesis.

```
> summary(lm(y ~ x, offset= 0.01*x))

Call:
lm(formula = y ~ x, offset = 0.01 * x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72384 -0.19138  0.06136  0.13320  0.69412 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  0.3229959  0.1521958   2.122   0.0472 *
x           -0.0014067  0.0009499  -1.481   0.1551  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.365 on 19 degrees of freedom
Multiple R-squared:  0.8116,	Adjusted R-squared:  0.8017 
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08
```

Harshith Bondada
M10724854

5. Using the output from summary (), suppose we want to test for null hypothesis of H_0: $\beta_1 = 0.02$ against the alternative hypothesis H_1: $\beta_1 \neq 0.02$, what do you conclude? Reject or not reject? Explain why.

Here, we reject the null hypothesis for B_1=0.02.

We can clearly see that the t test say's that the B_1 is -12.008 SD errors below the null hypothesis. When taken the absolute value we get 12.008, which is greater than 2. Hence, we reject the null hypothesis. The p value is way too small. i.e., less than 0.05.

```
> summary(lm(y ~ x, offset= 0.02*x))

Call:
lm(formula = y ~ x, offset = 0.02 * x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72384 -0.19138  0.06136  0.13320  0.69412 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.3229959  0.1521958   2.122   0.0472 *  
x           -0.0114067  0.0009499 -12.008 2.56e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.365 on 19 degrees of freedom
Multiple R-squared:  0.8116,	Adjusted R-squared:  0.8017 
F-statistic: 81.83 on 1 and 19 DF,  p-value: 2.579e-08
```

Harshith Bondada
M10724854

6. Repeat the same questions (1-3) for the date set <bus.csv>. Description: Cross-sectional analysis of 24 British bus companies (1951).

6.1. Read <bus.csv> into R Use response variable = Expenses per car mile (pence), covariate = Car miles per year (1000s).

We are reading the data of the bus(1).csv into homework3 and taking the response variable y as the Expenses per car mile and covariate x as the car miles.

```
homework3=read.csv("bus(1).csv")

attach(homework3)
y1=Expenses.per.car.mile..pence.
x1=Car.miles.per.year..1000s.
model2=lm(y1~x1, data=homework3)
plot(y1~x1)
abline(model2, lwd=2)
```

In the below output screenshot, we can interpret that the data is successfully read into the homework3 and a model2 has been generated.

```
> homework3=read.csv("bus(1).csv")
> attach(homework3)

> y1=Expenses.per.car.mile..pence.
> x1=Car.miles.per.year..1000s.
> model2=lm(y1~x1, data=homework3)
> plot(y1~x1)
> abline(model2, lwd=2)
```

6.2. Obtain R^2, explain what it means.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

The higher the R-squared, the better the model fits your data.

```
summary(model2)$r.square
summary(model2)
```

The summary of the model2 is shown in the below screenshot.

We get the R2 to be around 15%. This means that the 15% of the response variable can be explained by the linear regression model.

This means that the more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line.

Here, this means that the regression line is 15% good fit for the data.

```
> summary(model2)$r.square
[1] 0.1582641
> summary(model2)

Call:
lm(formula = y1 ~ x1, data = homework3)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0123 -0.9417 -0.1894  0.8993  2.6176 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.878e+01  4.075e-01  46.085   <2e-16 ***
x1          -4.450e-05  2.188e-05  -2.034   0.0542 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.347 on 22 degrees of freedom
Multiple R-squared:  0.1583,	Adjusted R-squared:   0.12 
F-statistic: 4.136 on 1 and 22 DF,  p-value: 0.0542
```

6.3. Perform the following hypothesis testing and interval estimation using lm() and other related R functions.

6.3.1. Perform t tests, obtain t statistics and p values, interpret the results, make a conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

Null Hypothesis:

If there is a significant **linear** relationship between the independent variable X and the dependent variable Y, the slope will not equal zero. The **null hypothesis** states that the

Harshith Bondada
M10724854

slope is equal to zero, and the **alternative hypothesis** states that the slope is not equal to zero.

We can know the values of the t test, it's statistics and p value by using the summary function.

Here, we find the t value for the slope to be negative 2.03383, which means that the slope is -2.03383 standard errors below 0, When we take the absolute value we get 2.03383 which is almost equal to 2. Hence, we do not reject the null hypothesis.

```
45  summary(model2)$coeff[,3]
```

```
> summary(model2)$coeff[,3]
(Intercept)          x1
   46.08506    -2.03383
```

We also get to know the P value from the summary function.

Similarly, with the slope it is greater than 0.05. Hence, we do not reject the null hypothesis.

```
> summary(model2)$r.square
[1] 0.1582641
> summary(model2)

Call:
lm(formula = y1 ~ x1, data = homework3)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0123 -0.9417 -0.1894  0.8993  2.6176 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.878e+01  4.075e-01  46.085   <2e-16 ***
x1          -4.450e-05  2.188e-05  -2.034   0.0542 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.347 on 22 degrees of freedom
Multiple R-squared:  0.1583,	Adjusted R-squared:   0.12 
F-statistic: 4.136 on 1 and 22 DF,  p-value: 0.0542
```

6.3.2. Perform ANOVA test (F test), obtain F statistic and p value, interpret the results, make conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

Null hypothesis: The fit of the intercept-only model and your model are equal.

We can find the F-statistic and the p value using the summary function.

```
summary(model2)
```

The values of the F-statistic and p value are found at the bottom of the summary.

We find the ratio for the variances is 4.136 which is small F value.

When there is a large F value, the corresponding P value is small.

Here, the p value is greater than 0.05, we reject the Null hypothesis.

```
Residual standard error: 1.347 on 22 degrees of freedom
Multiple R-squared:  0.1583,    Adjusted R-squared:   0.12
F-statistic: 4.136 on 1 and 22 DF,  p-value: 0.0542
```

We do not reject the null hypothesis.

6.3.3. Compute confidence interval for coefficients, fitted values (mean response), interpret the meanings of these quantities.

We can compute the **confidence interval** for the fitted values using the **predict()** function.

```
head(y1)
head(x1)
d1=data.frame(y1,x1)
head(d1)
predict.lm(model2, interval="confidence")
```

I am first creating data frame for y1 and x1.

And then generating the head of the data frame for explaining it easily.

You can see in the below screenshot that I have used the predict function with two arguments. The first one takes our generated model and the second one takes the type of interval. Since, we want confidence, we mention it’s a confidence interval.

Harshith Bondada
M10724854

```
> head(d)
     y  x
1 0.27 12
2 0.14 20
3 0.33 20
4 0.81 46
5 0.84 48
6 1.08 92
```

This means that for a observation of y1 and x1

The point estimate is between the lwr and upr values.

Let me explain, with a simple example. Let us take the first observation.

This means that for y=19.76 with regressor x= 6235, the point y^ estimate is 18.50435 in the interval range [17.83996, 19.16874] with 95% probability.

Harshith Bondada
M10724854

```
> head(y1)
[1] 19.76 17.85 19.96 16.80 18.20 16.71
> head(x1)
[1]  6235 46230  7360 28715 21934  1337
> d1=data.frame(y1,x1)
> head(d1)
     y1    x1
1 19.76  6235
2 17.85 46230
3 19.96  7360
4 16.80 28715
5 18.20 21934
6 16.71  1337
> predict.lm(model2, interval="confidence")
        fit      lwr      upr
1  18.50435 17.83996 19.16874
2  16.72461 15.14429 18.30493
3  18.45429 17.81459 19.09398
4  17.50401 16.61724 18.39078
5  17.80576 17.12523 18.48629
6  18.72231 17.92084 19.52377
7  17.98611 17.38583 18.58639
8  18.67861 17.90780 19.44942
9  17.97904 17.37647 18.58161
10 18.73076 17.92321 19.53831
11 18.68497 17.90978 19.46016
12 18.19143 17.62077 18.76209
13 18.62245 17.88895 19.35595
14 18.10969 17.53614 18.68324
15 16.68994 15.07660 18.30328
16 18.33063 17.73733 18.92392
17 18.50827 17.84182 19.17471
18 17.75436 17.04387 18.46486
19 17.86734 17.21895 18.51574
20 18.36129 17.75861 18.96396
21 18.73606 17.92468 19.54744
22 18.61057 17.88462 19.33651
23 18.08512 17.50835 18.66190
24 18.43805 17.80568 19.07041
> 
```

We can compute the **confidence interval** for the coefficients using the **confint( )** function.

This interval contains the true parameter with 95% probability.

Interval estimate [1.793660e+01 1.962700e+01]

This contains the true intercept 95% probability.

Slope estimate [-8.987441e-05 8.761294e-07]. This contains the true slope With 95% probability.

```
confint(model2)
```

```
> confint(model2)
                    2.5 %       97.5 %
(Intercept)  1.793660e+01 1.962700e+01
x1          -8.987441e-05 8.761294e-07
```

6.3.4. Plot data points, the regression line, the confidence interval for fitted values (to show that the interval is wider on both sides and narrow in the center).

We can obtain the confidence intervals using the lines function.

We first create a new variable and specify it's range.

Since, our range is beyond 45000 on the x-axis, we specify the range in newx to be around (0, 49000).

```
newx1<-seq(0,49000)
conf<-predict(model2,newdata=data.frame(x1=newx1),interval = c("confidence"),level = 0.95,type="response")
plot(x1,y1,pch=20)
model2 <- lm(y1 ~ x1, data=homework3)
abline(model2,col="blue")
lines(newx1,conf[,2],col="red",lty=100)
lines(newx1,conf[,3],col="red",lty=100)
```

The below screenshot describes the regression line, and the confidence interval

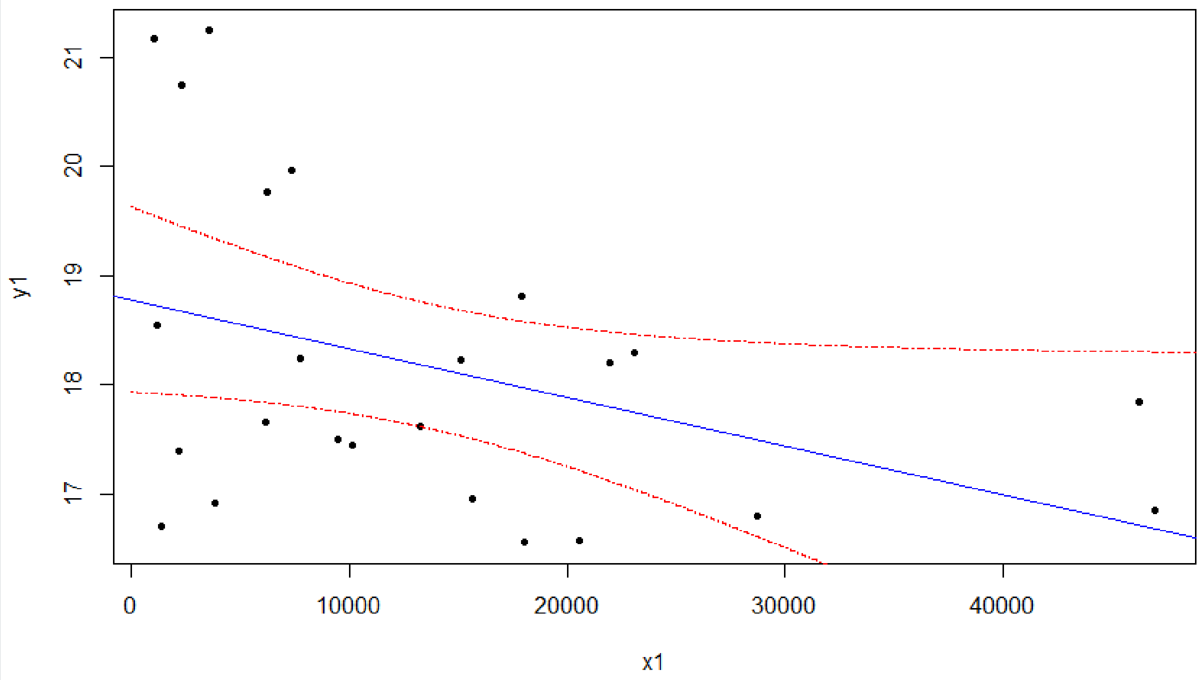
Lines for the values.

The blue line is the regression line. The other two red lines are the upper bound line and the lower bound line.

The interval estimation for the mean response of the first observation is [18,19.75].

In other words, [18,19.75] contains the true mean response with 95% probability.

Harshith Bondada
M10724854



6.4. Using the output from summary (), suppose we want to test for null hypothesis of $H_0: \beta_1 = 0.01$ against the alternative hypothesis $H_1: \beta_1 \neq 0.01$, what do you conclude? Reject or not reject? Explain why.

Here, we reject the null hypothesis for B_1=0.01.

We can clearly see that the t test say's that the B_1 is -459.08 SD errors below the null hypothesis. When taken the absolute value we get 459.08, which is greater than 2. Hence, we reject the null hypothesis. The p value is way too small. i.e., less than 0.05.

Harshith Bondada
M10724854

```
> summary(lm(y1 ~ x1, offset= 0.01*x1))

Call:
lm(formula = y1 ~ x1, offset = 0.01 * x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0123 -0.9417 -0.1894  0.8993  2.6176 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.878e+01  4.075e-01   46.09   <2e-16 ***
x1          -1.004e-02  2.188e-05 -459.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.347 on 22 degrees of freedom
Multiple R-squared:  0.1583,	Adjusted R-squared:   0.12 
F-statistic: 4.136 on 1 and 22 DF,  p-value: 0.0542
```

6.5. Using the output from summary(), suppose we want to test for null hypothesis of $H_0: \beta_1 = 0.02$ against the alternative hypothesis $H_1: \beta_1 \neq 0.02$, what do you conclude? Reject or not reject? Explain why.

Here, we reject the null hypothesis for B_1=0.02.

We can clearly see that the t test say's that the B_1 is -916.13 SD errors below the null hypothesis. When taken the absolute value we get 916.13, which is greater than 2. Hence, we reject the null hypothesis. The p value is way too small. i.e., less than 0.05.

Harshith Bondada
M10724854

```
> summary(lm(y1 ~ x1, offset= 0.02*x1))

Call:
lm(formula = y1 ~ x1, offset = 0.02 * x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0123 -0.9417 -0.1894  0.8993  2.6176 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.878e+01  4.075e-01   46.09   <2e-16 ***
x1          -2.004e-02  2.188e-05 -916.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.347 on 22 degrees of freedom
Multiple R-squared:  0.1583,	Adjusted R-squared:   0.12 
F-statistic: 4.136 on 1 and 22 DF,  p-value: 0.0542
```